Development of machine learning models to improve ESG scores

Mingze Li

Environmental, Social, and Governance (ESG) factors are crucial in the investment decision of a company, considering its sustainability and societal impact.

We want to use natural language processing (NLP) to extract topics and transfer the unstructured data to structured data that can be used for analyzing the text in terms of ESG concepts.

My role in the project is to develop and deploy python scripts to visualize topics and the result of the ESG analysis.

1.Topic modeling

Topic modeling is a type of Natural Language Processing (NLP) task that utilizes unsupervised learning methods to extract out the main topics of some text data we deal with. Instead of pre-training on the data that have associated topic labels, the algorithms try to discover the underlying patterns, in this case, the topics, directly from the data itself.

Our goal is to extract the topics from unlabeled reports. **Here we build our topics_modeling with the help of LDA, BERTopic and NMF**.
This extractor has a built-in function:
`extract_document_topics(path,num_topics)`.
When the function is called, it will read the data in the path, train the three models mentioned above and write their summarized topics into a txt file.
At the same time, we write the function
`topic_classification(self, txt)` for each nlp model so that we can get the most similar topic after training these models.

our input data:

	page_num	type	check	text
0	1	header	OK	Building Tomorrow Together
1	1	header	OK	2021 Environmental, Social and Governance Report
2	2	header	OK	TD Bank Group 2021 ESG Report
3	2	header	OK	Table of Contents
4	2	header	OK	Introduction
5	2	header	OK	Performance Highlights for Investors
6	2	header	OK	About This Report
7	2	header	OK	1.1 A Message From Our Leadership
8	2	header	OK	1.2 About TD
9	2	header	OK	1.3 Implementing Our ESG Strategy
10	2	header	OK	1.4 Global Developments Shaping Our Future
11	2	header	OK	1.5 ESG Trends
12	2	header	OK	1.6 How We Listen to Stakeholders
13	2	header	OK	1.7 Our ESG Material Topics
14	2	header	OK	1.8 ESG Scorecard and Goals
15	2	header	OK	Governance
16	2	header	OK	2.1 Corporate Governance and Integrity
17	2	header	OK	2.2 Risk Management
18	2	header	OK	2.3 Data Security and Privacy
19	2	header	OK	2.4 Human Rights
20	2	header	OK	2.5 Tax
21	2	header	OK	Environmental
22	2	paragraph	check	3.1 Our E Journey: A Message From Our Head of Environment
23	2	header	OK	3.2 Climate Change
24	2	header	OK	3.3 Sustainable Finance
25	2	header	OK	3.4 Lending
26	2	header	OK	3.5 Investing
27	2	header	OK	3.6 Responsible Resource Use
28	2	header	OK	Social
29	2	paragraph	check	4.1 Our S Journey: A Message From Our Head of U.S. Corporate Citizenship
30	2	header	OK	4.2 Financial and Economic Inclusion
31	2	header	OK	4.3 Economic Value
32	2	header	OK	4.4 Social Inclusion
33	2	header	OK	4.5 Volunteerism
34	2	header	OK	4.6 Responsible Sourcing
35	2	header	OK	4.7 Customer Experience
36	2	header	OK	4.8 Product and Service Responsibility
37	2	header	OK	4.9 Diversity and Inclusion
38	2	header	OK	4.10 Talent Attraction, Development and Retention
39	2	header	OK	4.11 Health and Well being
40	2	header	OK	2021 Awards and Recognition 83

our output file:

```
when num_topics = 14:

LDA
['risk,climate,tax,table,change,management,asset,portfolio,assessment,tds,investment,process,business,impact,credit',
 'retention,attraction,talent,bond,development,issuance,gas,greenhouse,carbon,ghg,emission,proceeds,outcome,sustainability,community',
 'report,esg,group,bank,progress,experience,matter,customer,action,subsidiary,plan,target,goal,addition',
 'community,colleague,future,minority,woman,workplace,work,development,lgbtq2,director,help,people',
 'loan,business,account,canada,online,credit,program,level,employee,facility,customer,mobile,president,app',
 'diversity,inclusion,year,corner,analyst,unit,result,vehicle,number,hour,position',
 'target,responsibility,service,product,emission,executive,net,reduction,cyber,scope,baseline,representation',
 'introduction,finance,leadership,message,support,colleague,program,learning,expertise,head,solution,development,capability,business,cybersecurity,housing',
 'index,award,topic,volunteerism,disclosure,employer,period,accounting,anti,year',
 'governance,esg,inclusion,value,trend,strategy,integrity,sustainability,proxy,goal,system,time',
 'journey,emission,scope,use,resource,manager,location,committee,tonne,intensity,approach,heating',
 'statement,caution,privacy,security,conduct,training,code,employee,law,compliance,policy,data',
 'customer,complaint,lending,service,product,advice,insurance,consumer,survey,practice,office,process,line,canada',
 'performance,data,health,provision,investor,loss,information,october,asset,business,owner']

NMF:
['matter,goal,scorecard,topic,material,trend,strategy,bank,td,group',
 'tdam,unit,page,progress,target,result,year,privacy,security,investor',
 'retention,attraction,diligence,journey,bond,sustainability,risk,performance,inclusion',
 'security,board,journey,management,tax,committee,performance,strategy,model,policy',
 'carbon,process,management,exposure,diligence,emission,journey,finance,lending,change',
 'topic,message,development,trend,strategy,introduction',
 'world,score,october,engagement,employee,year,progress,unit,target,accounting',
 'key,mining,forestry,list,report,acronym,endnotes',
 'material,indirect,subsidiary,shareholder,page,human,risk,bank,statement,caution',
 'island,territory,provide,information,key,list,report,page,acronym',
 'document,index,fn,loan,morningstar,post,aum,result,ar,employee',
 'facility,minority,year,unit,target,result,award,recognition',
 'community,education,value,diversity,inclusion',
 'support,change,management,community,colleague,development,responsibility,climate,business,product']

BERT:
['risk,customer,colleague,business,bank,climate',
 'year,fn,number,donation,amount,loan',
 'emission,ghg,co,carbon,climate,target,energy',
 'community,woman,diversity,minority,representation,canada,inclusion',
 'esg,strategy,expertise,centre,business,governance,report,stakeholder',
 'level,management,executive,manager,men,vice,president',
 'privacy,cybersecurity,security,threat,information,risk,td,technology,protect',
 'tdi,insurance,weather,disaster,resilience,damage,panel,climate',
 'tdam,fund,investment,esg,asset,investor',
 'tax,property,payroll,sale,income,capital,behalf,jurisdiction,insurance',
 'subsidiary,material,law,security,bank,disclosure,authority',
 'bond,sustainability,issuance,proceeds',
 'yes,,']
```

Latent Dirichlet Allocation (LDA):

The basic assumption for LDA is that each of the documents can be represented by the distribution of topics which in turn can be represented by some word distribution.
LDA can be used to extract main topics from the text, a pre-trained LDA model can also classify the unknown text into the extracted topics.

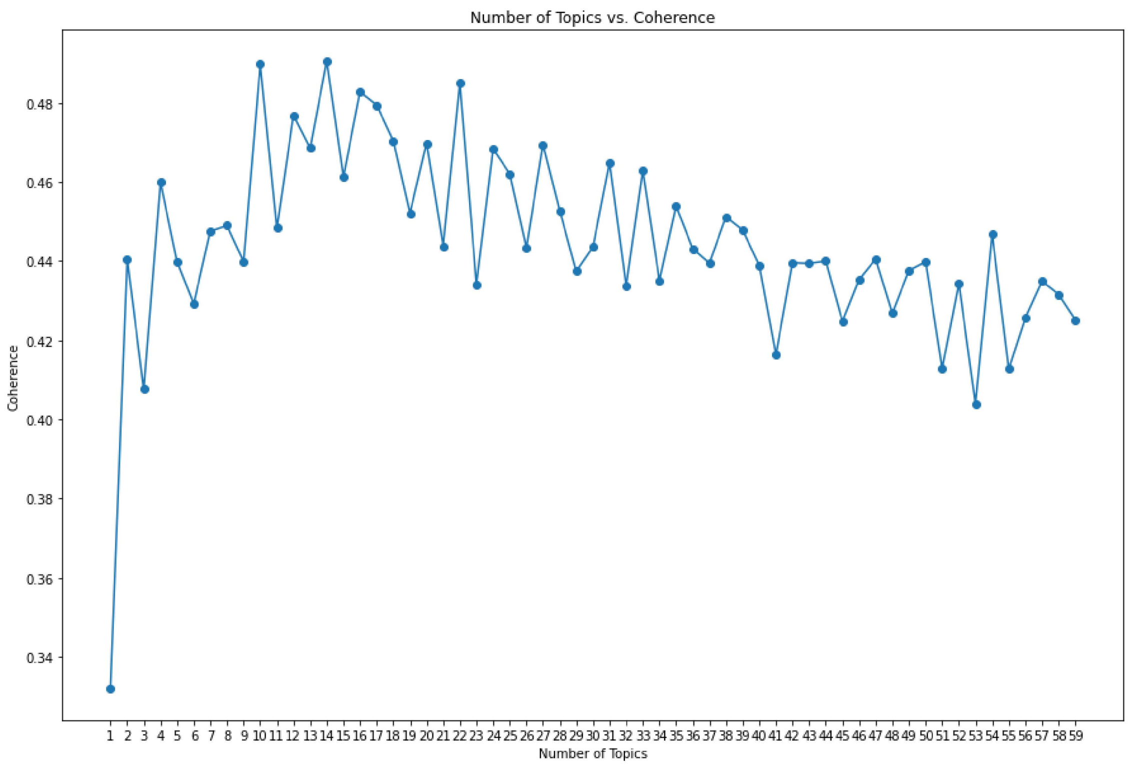
Detect the num_topics:

Except for BERTopic, all other models require us to set the number of topics manually. Therefore we use a loop to visualize the effect of different numbers of topics on the coherence score of the LDA model. Since we wanted to extract as many keywords as possible, we tested the score from 1 to 60.

Coherence score was applied to evaluate the association of the extracted topics so that we can choose a suitable number of topics. C_v is one of the more widely used approaches, normally somewhere between 0.5 and 0.7 would be considered as a decent score.

```
# select the num_topics with the highest coherence score
topics = []
score = []
for i in range(1, num_max, 1):
    lda = gensim.models.LdaMulticore(
        corpus=self.bow_corpus,
        id2word=self.dic,
        iterations=10,
        num_topics=i,
        workers=3,
        passes=10,
        random_state=42,
    )
    cm = CoherenceModel(model=lda, corpus=self.bow_corpus,
texts=self.corpus, coherence="c_v")
    topics.append(i)  # Append number of topics modeled
    score.append(cm.get_coherence())  # Append coherence scores to list
    print("when num_topics = " + str(i))
    print("c_v score: " + str(cm.get_coherence()))
n_topics = topics[score.index(max(score))]
```

```
plt.figure(figsize=(15, 10))
plt.plot(topics, score)
plt.scatter(topics, score)
plt.title("Number of Topics vs. Coherence")
plt.xlabel("Number of Topics")
plt.ylabel("Coherence")
plt.xticks(topics)
plt.show()
```



We can see that num_topics = 12 and num_topics = 14 are two peaks of the highest score. Here we choose num_topics=14 to get more keywords.

pyLDAvis:

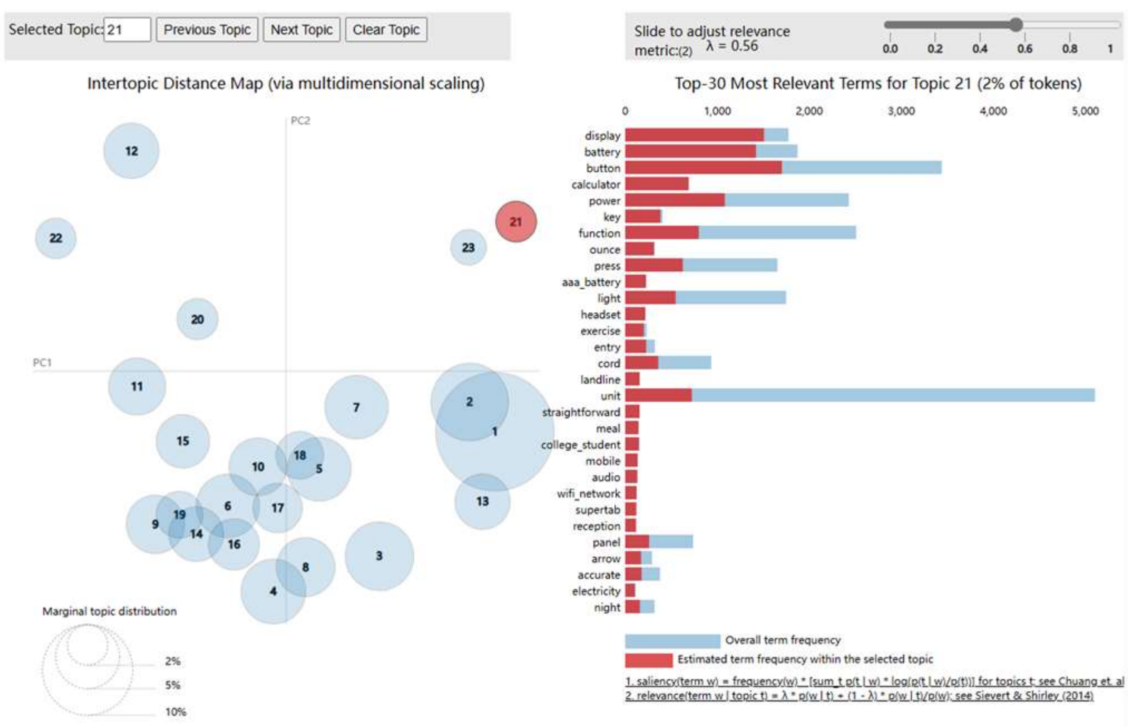
pyLDAvis is designed to help users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization.

```
pyLDAvis.enable_notebook()

bow_corpus = [lda.dic.doc2bow(doc) for doc in lda.corpus]

vis = pyLDAvis.gensim_models.prepare(lda.model, bow_corpus, lda.dic,
sort_topics=False)

pyLDAvis.display(vis)
```

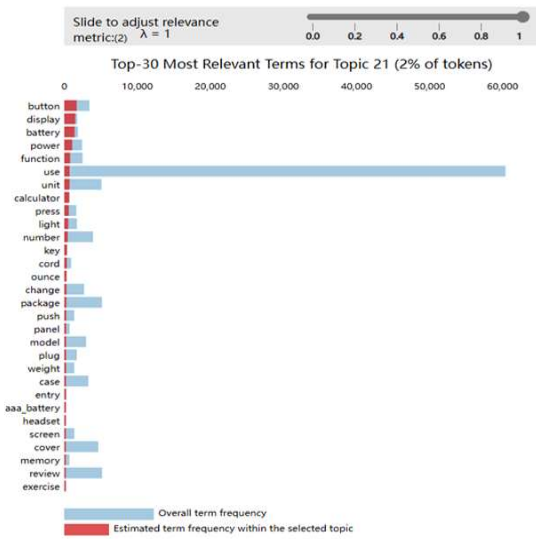


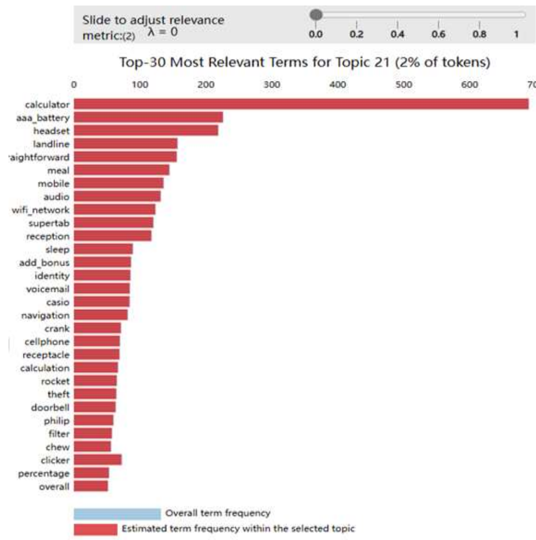
Relevance metric λ:

As we can see, the parameter λ is adjustable.

relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w)

When λ = 1, the words are sorted in order of absolute number of times, when λ = 0, they are sorted in order of proportion p(w | t)/p(w).





Therefore, we can adjust the value of the λ to extract the top words we want. In our extractor, we find λ = 0.6 is a suitable value that can take both cases into account.

For all the models, we only keep the nouns when generating the topics because they are more likely to be considered topics.

```
LDA_topics = []

num_terms = 20 # Adjust number of words to represent each topic

lambd = 0.6 # Adjust this accordingly based on tuning above

for i in range(1,num_topics+1): #Adjust this to reflect number of topics
chosen for final LDA model

    topic = vis.topic_info[vis.topic_info.Category ==
'Topic'+str(i)].copy()

    topic['relevance'] = topic['loglift']*(1-lambd)+topic['logprob']*lambd

    list_topics = topic.sort_values(by='relevance',
ascending=False).Term[:num_terms].values

    is_noun = lambda pos: pos[:2] == 'NN'

    all_mouns = ','.join([word for (word, pos) in nltk.pos_tag(list_topics)
if is_noun(pos)])

    LDA_topics.append(all_mouns)
```

Classification of unknown text:

A pre-trained LDA model can also be used to classify unknown text. after transform the text into bow_vector, we use lda_model.get_document_topics(bow_vector) to get the index of possible topics, we select the one with maximum possibility

```
def para_preprocess(text):
    result = []
    text = data_cleaning(text)
    words = [w for w in nltk.tokenize.word_tokenize(text) if (w not in
stopwords)]
    # word_tokenize function tokenizes text on each word by default
    words = [lem.lemmatize(w) for w in words if len(w) > 2]
    result.append(words)
    return result
txt = The feedback we received reinforced our belief that Cheggs
mission and values are critical to our business success and are deeply
integrated into our culture and processes.

bow_corpus = para_preprocess(txt)
bow_vector = self.dic.doc2bow(bow_corpus[0])  # transform the corpus to
doc2bow
# print(self.model.get_document_topics(bow_vector))
index=max(self.model.get_document_topics(bow_vector),key=itemgetter(1))
[0]  # find the most similair topic
```

BERTopic:

BERTopic is a topic modeling python library that combines transformer embeddings and clustering model algorithms to identify topics in NLP. Because the embedding vectors usually have very high dimensions, dimension reduction techniques are used to reduce the dimensionalities. The default algorithm for dimension reduction is UMAP (Uniform Manifold Approximation & Projection). Compared with other dimension reduction techniques such as PCA (Principal Component Analysis), UMAP maintains the data's local and global structure when reducing the dimensionality, which is important for representing the semantics of the text data.

Unlike LDA, BERTopic will automatically determine the number of topics based on the training data. If the detected number is larger than we expect, we can reduce the number of topics. But in the opposite case, we can't do anything about it

```
data = data[data["type"] == "paragraph"]
data.reset_index(drop=True, inplace=True)
self.doc = data_lem_without_tokenize(data)
# Initiate UMAP
umap_model = UMAP(n_neighbors=15, n_components=5, min_dist=0.0,
metric="cosine", random_state=100)
# Initiate BERTopic
self.topic_model = BERTopic(umap_model=umap_model, language="english",
calculate_probabilities=True)

# Run BERTopic model
topics, probabilities = self.topic_model.fit_transform(self.doc)
if len(topics) > num_topics:
    self.topic_model.reduce_topics(self.doc, nr_topics=num_topics)
print("BERT model trained")
```

Classification of unknown text:

BERTopic can also classify the topic of unknown text. It is easier because we don't need to convert unknown text to a corpus vector.

```
new_review = "The feedback we received reinforced our belief that
Cheggs mission and values are critical to our business success and are
deeply integrated into our culture and processes."
# Find topics
num_of_topics = 2
similar_topics, similarity = self.topic_model.find_topics(new_review,
top_n=num_of_topics);
# Print results
print(fThe top {num_of_topics} similar topics are {similar_topics}, and
the similarities are {np.round(similarity,2)})
# Print the top keywords for the top similar topics
for i in range(num_of_topics):
  print(fThe top keywords for topic {similar_topics[i]} are:)
  print(self.topic_model.get_topic(similar_topics[i]))
```

- *The top 2 similar topics are [4, 3], and the similarities are [0.28 0.27]*
- *The top keywords for topic 4 are:*
- *[(highlights, 0.33332467418214173), (framework, 0.32994210545593106), (awards, 0.32667194273220235), (overview, 0.3252125540833023), (sdgs, 0.3174695866613478), (pillars, 0.3154073929828161), (materiality, 0.31458557196024445), (oversight, 0.30906292061057355), (disclosures, 0.2988858404806048), (capital, 0.021435091302897225)]*
- *The top keywords for topic 3 are:*
- *[(esg, 0.360531237433342), (disclosures, 0.09900062445779519), (assessment, 0.06670067078813456), (key, 0.06451340029147333), (report, 0.059467890023657786), (see, 0.05812406712027032), (materiality, 0.052100440789361956), (students, 0.05201799042720433), (stakeholders, 0.05196884144122686), (formal, 0.05196884144122686)]*

Non-Negative Matrix Factorization (NMF)

Non-negative matrix factorization is also a non-supervised learning technique which performs clustering as well as dimensionality reduction. It can be used in combination with the TF-IDF scheme to perform topic modeling.

```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_vect = TfidfVectorizer(max_df=0.8, min_df=2, stop_words=english)
doc_term_matrix = tfidf_vect.fit_transform(df_input_lem.values.astype(U))
from sklearn.decomposition import NMF

nmf = NMF(n_components=8, random_state=42)
nmf.fit(doc_term_matrix)
for i,topic in enumerate(nmf.components_):
    print(fTop 10 words for topic #{i}:)
    print([tfidf_vect.get_feature_names_out()[i] for i in 
topic.argsort()[-10:]])
    print(\n)
```

For all paragraphs as the training data, NMF can classify their topics.

```
topic_values = nmf.transform(doc_term_matrix)
df_para[Topic] = topic_values.argmax(axis=1)
df_para.head(20)
```

	Unnamed: 0	page_number	type	text	Topic
0	3	2	paragraph	Over the last decade we have focused on puttin...	0
1	4	2	paragraph	The pace of evolution within the learning ecos...	0
2	6	2	paragraph	Weve also seen the challenges of the last few ...	0
3	7	2	paragraph	Chegg seeks to always serve as a valued and re...	0
4	9	3	paragraph	We know that the heart of Chegg is our incredi...	5
5	10	3	paragraph	In 2021, Chegg collectively donated 1,400,000 ...	0
6	11	3	paragraph	The challenges of the last few years have had ...	6
7	12	3	paragraph	We are a mission driven company with an enormo...	0
8	13	3	paragraph	Sincerely,	0
9	14	3	paragraph	Dan Rosensweig CEO, President, and Co Chairper...	2
10	15	4	paragraph	When it comes to our most valuable resource, o...	5

2. keywords visualization

visualization can be a powerful tool for helping users understand keywords, it presents information in a clear and concise way, making it easier for users to understand and absorb. Instead of being presented with a list of top words, users can see the relationships between different keywords and how they relate to the larger topic or concept. Therefore, **we create a keyword_extractor to help the user apply different visualizations on the data.**

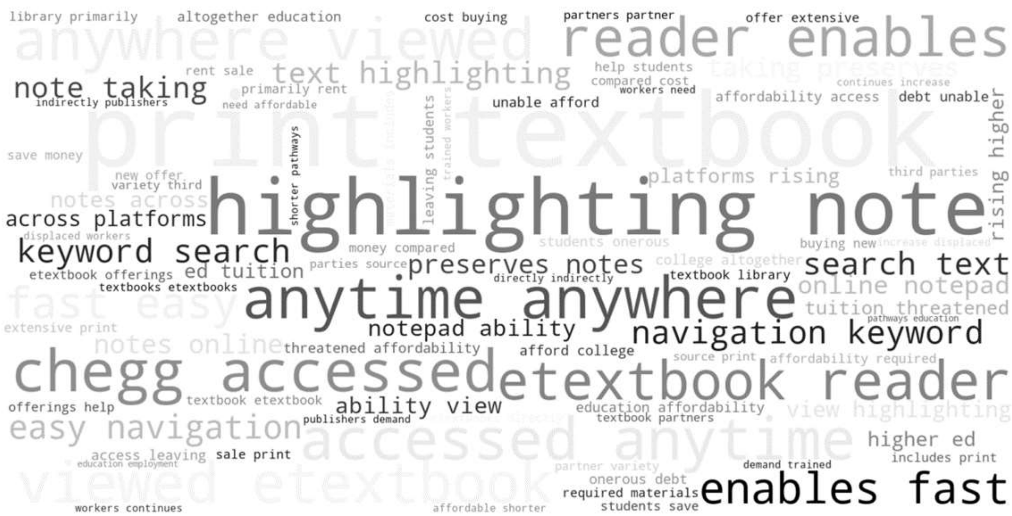
our input text (already pre-processed):

```
text = """
from  Chegg can be accessed anytime and anywhere and are
viewed through our eTextbook reader that enables fast and
easy  navigation, keyword search, text highlighting, note taking
and further preserves those notes in an online notepad
with the ability  to view highlighting and notes across platforms.
Rising Higher Ed tuition has threatened
affordability and access, leaving many students  with onerous
debt or unable to afford college  altogether.
Education Affordability  Required Materials includes our print
textbook and eTextbook offerings, which help students
save money compared to  the cost of buying new. We offer an
extensive print textbook library primarily for rent and
also for sale both on our own and  through our print textbook
partners. We partner with a variety of third parties
to source print textbooks and eTextbooks directly  or indirectly
from publishers.  Demand for trained workers
continues to  increase but displaced workers need more
affordable and shorter pathways from education  to
employment.
"""
```

Word cloud

Our basic word cloud can display high frequency bigrams phrases.

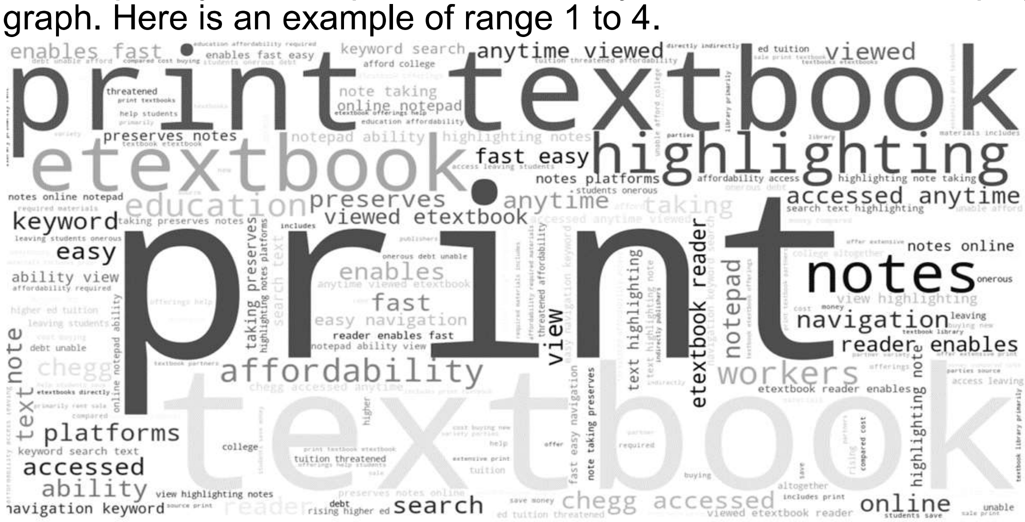
```
wordcloud_text =
WordCloud(collocation_threshold=2,collocations=True,background_color="white",colormap='binary', width=1600, height=800).generate(text)
plt.imshow(wordcloud_text, interpolation='bilinear')
plt.axis("off")
plt.show()
```



N-grams word cloud:

We also support n-grams word cloud. The users can define their desired n-grams range.

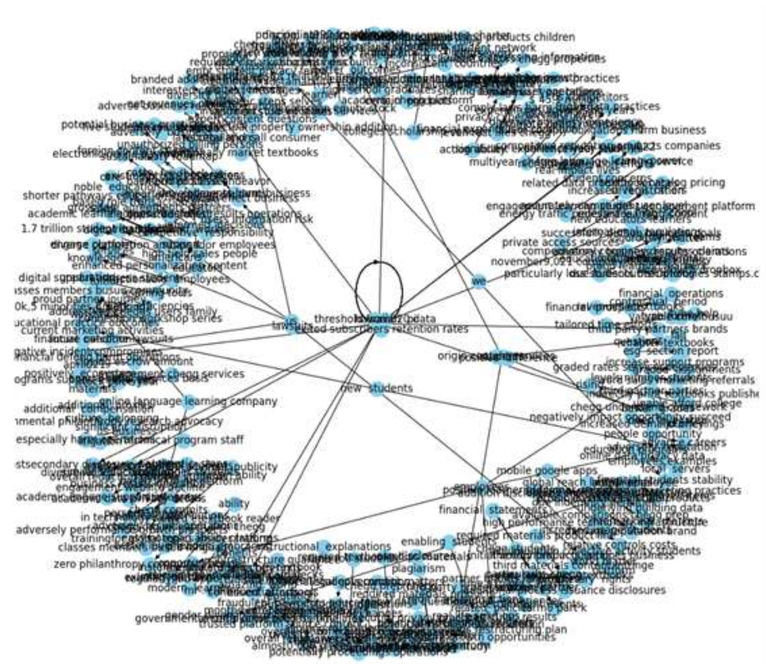
We use CountVectorizer to convert the input text into vectors, calculate the frequency of each phrase, and finally use word cloud to display the graph. Here is an example of range 1 to 4.

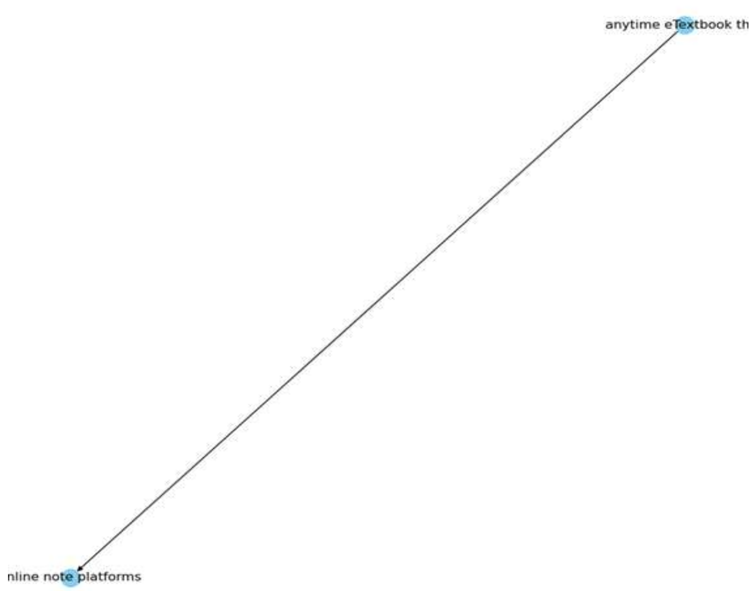


Relation plot:

We sentence-tokenize the input text, analyze the relational words, subject and object in each sentence. Each line segment indicates that the two connected nouns are linked by a relational word. The users can define the relation words they desire, the default plot will be a diagram including all the relation words.

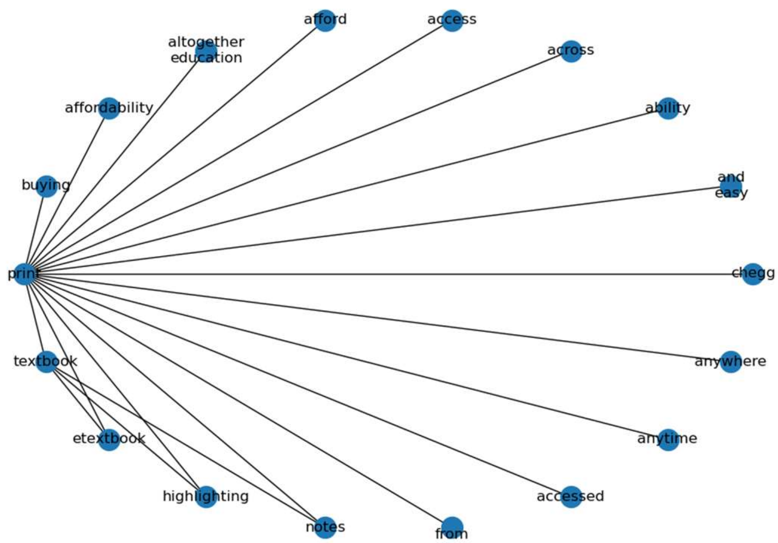
The above plot is without defining the relation words, the following is when relation = “accessed”





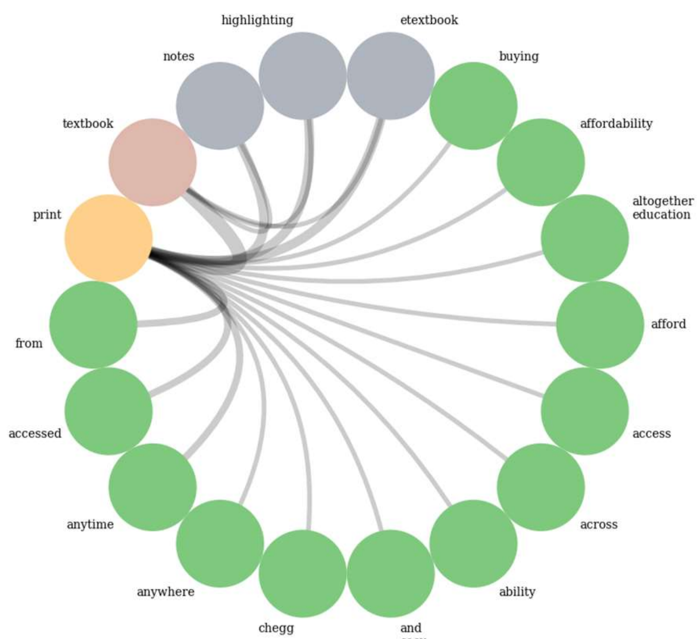
Shell diagram:

Shell diagram connects all related entities. (Lines are not weighted)



Circosplot:

Like the shell diagram, the circosplot connects also the related entities. But its connected lines are weighted. The thicker the line is, the stronger their relationship is (the more frequently they appear in a same sentence)

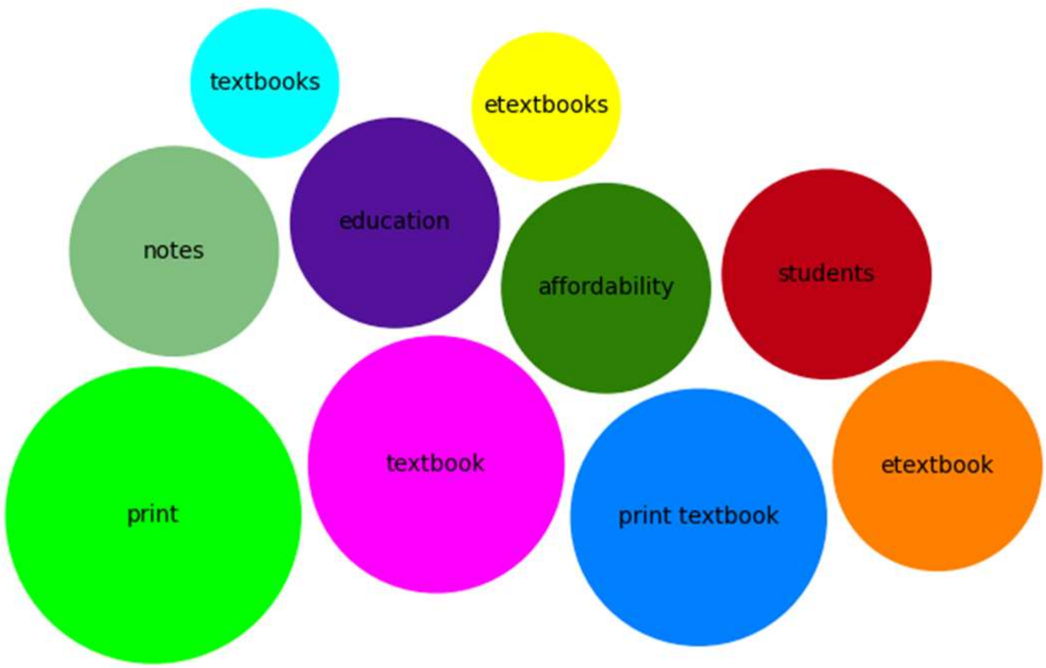


<u>packed bubble chart:</u>

A packed bubble chart displays 10 top words in a cluster of circles. It has three modes: mix, frequency and keybert.

In frequency mode, some of the top words appear frequently, but these words have no meaning and cannot be considered as keywords; in keybert mode, all the top words have the highest keybert score, but some of them only appear once or twice and do not represent the text very well.

As the most powerful mode, mix will combine the best of both. It will first get the top 30 words based on their keybert scores, then sort with their frequencies. This allows us to filter out the words that appear frequently but have no contribution to the meaning.

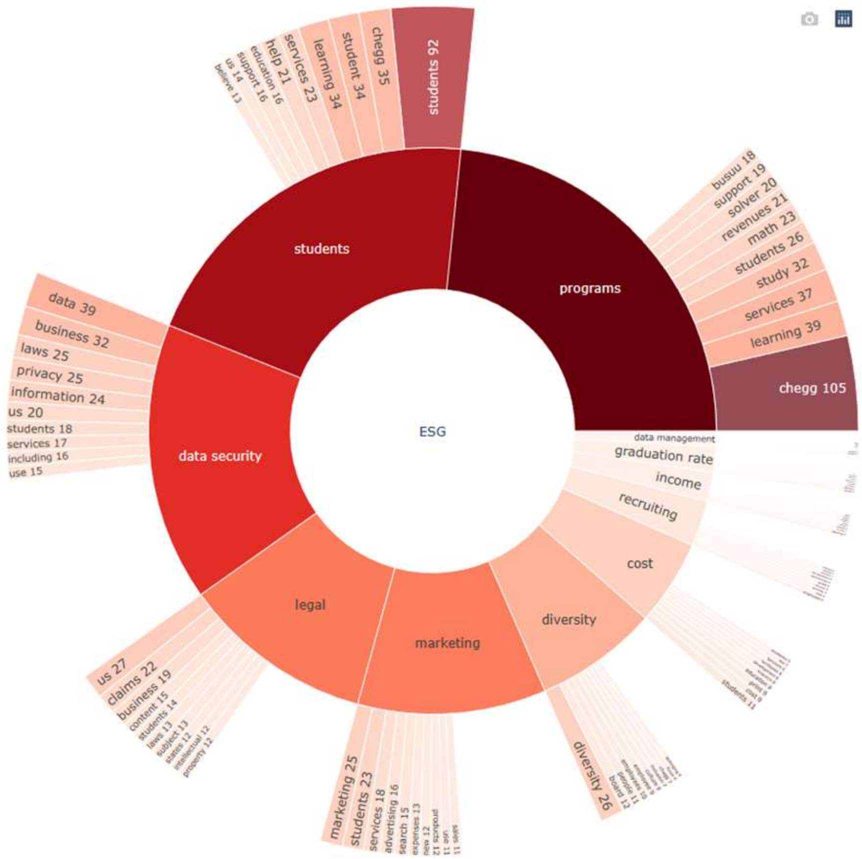


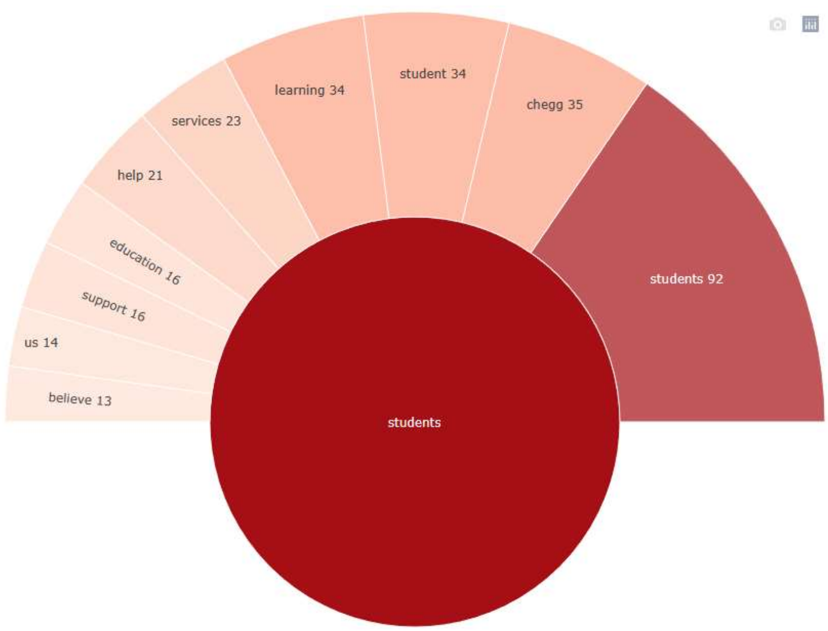
Interactive visualizations

We also have data in which the paragraphs are manually labeled with the corresponding topics. In that case, the interactive visualizations allow users to explore keywords in a dynamic way. It allows users to click on a topic and see related keywords, making it easier for them to understand the differences between each topic.

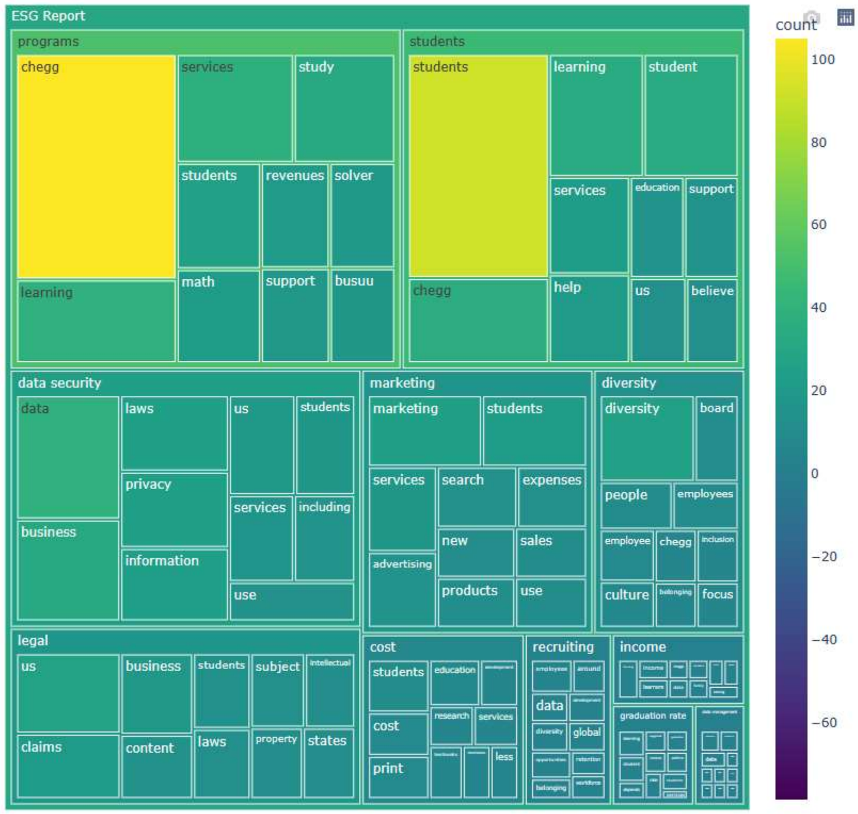
	A	B
1	Paragraph	topic
2	Support a modernized student financial aid model Address the key drivers of student debt 20% of Cheggs	income
3	Chegg promotes culture, belonging, and diversity (CB&D) across all levels, reflecting the multi-cultural learn	diversity
4	Student Graph. Our Student Graph is the accumulation of the collective activity of students in our learning p	students
5	The pace of evolution within the learning ecosystem has accelerated in recent years as students and educa	income
6	Our Chegg Services product line for students primarily includes Chegg Study, Chegg Writing, Chegg Math S	programs
7	27% first generation in college 20% family income of <$20k 55% minorities 62% female 23% >25yrs old 11% wo	diversity
8	CheggsAmbassador Program is designed to equip and empower every global employee to instill an inclu	diversity
9	Cheggs Leading with Inclusion Workshop helps define Diversity & Inclusion at Chegg, introduces employee	diversity
10	Launching our Chegg Ambassador Program, designed to equip and empower every employee to instill ar	diversity
11	We continue to increase our focus on diversity and inclusion within our recruiting and retention efforts	recruiting
12	More Help, Less Cost: The rising cost of education is an increasing impediment for aspiring learners all ove	cost
13	We know that the heart of Chegg is our incredible employees. They are one of our biggest competitive adv	diversity
14	Invest in workforce programs providing access to relevant skills training Support workforce programs deve	programs
15	If we do not attract more students or if students do not increase their level of engagement with our platfor	marketing
16	We are extremely proud of the fact that 94% of global employees believe people of all cultures and backgr	diversity
17	As of December 31, 2021, we had 1,736 employees, of which 1,613 were full-time and 123 were part-time.	recruiting
18	In addition to language learning, Chegg is expanding its Total Addressable Market by serving people acro	programs
19	Laws or regulations may be enacted which restrict or prohibit use of emails or similar marketing activities th	marketing
20	The regulatory framework for privacy issues worldwide is currently in flux and is likely to remain so for the fo	data security
21	Students are more diverse in age, race, and income than previous generations. Embracing technology can	cost
22	We have internal and publicly posted policies regarding our collection, processing, use, disclosure, deletion	data security
23	In the ordinary course of business, we collect, process, store, and use personal information and data supplie	data security
24	We believe that adhering to our core value of putting students first is essential to our success and in the bes	students
25	The Family Educational Rights and Privacy Act (FERPA) protects the privacy of student records and gives stu	data security
26	In 2021 Chegg broadened its reach to more students, especially internationally, and we took another step	programs
27	Moreover, as the education industry continues to evolve, increasing regulation by federal, state, and foreigr	data security
28	Students and their families lack reliable, timely, and accurate information to make informed decisions abou	cost
29	Help students achieve better outcomes	students
30	. We strive to improve the overall return on investment in education by helping people learn more in less ti	cost
31	Chegg Study Pack. Our Chegg Study Pack is a premium subscription bundle including our Chegg Study, Che	programs
32	We strive to improve the overall return on investment in education by helping students learn more in less ti	cost
33	On December 22, 2021, Steven Leventhal, individually and on behalf of all others similarly situated, filed a pu	legal
34	Chegg is a mission-driven company. We put learners first and seek to improve their outcomes in school anc	cost
35	Students say that Chegg helps them better understand the concepts they are studying in school Students s	students
36	We use several major direct marketing channels to reach students. We deploy search engine optimization (S	marketing
37	Our ability to attract and retain students and increase their engagement with our learning platform depend	marketing
38	Students subscribe to our subscription services, collectively referred to as our Chegg Services, which can be	programs
39	the types of classes our students are taking and whether they choose to take those classes pass/fail;	students
40	We maintain personal data regarding students, tutors, and educators, including names and, in many cases, i	data security
41	We aim to support and accelerate the path students take from learning to earning. This includes online toc	students
42	We aim to support and accelerate the path students take from learning to earning. This includes online too	students

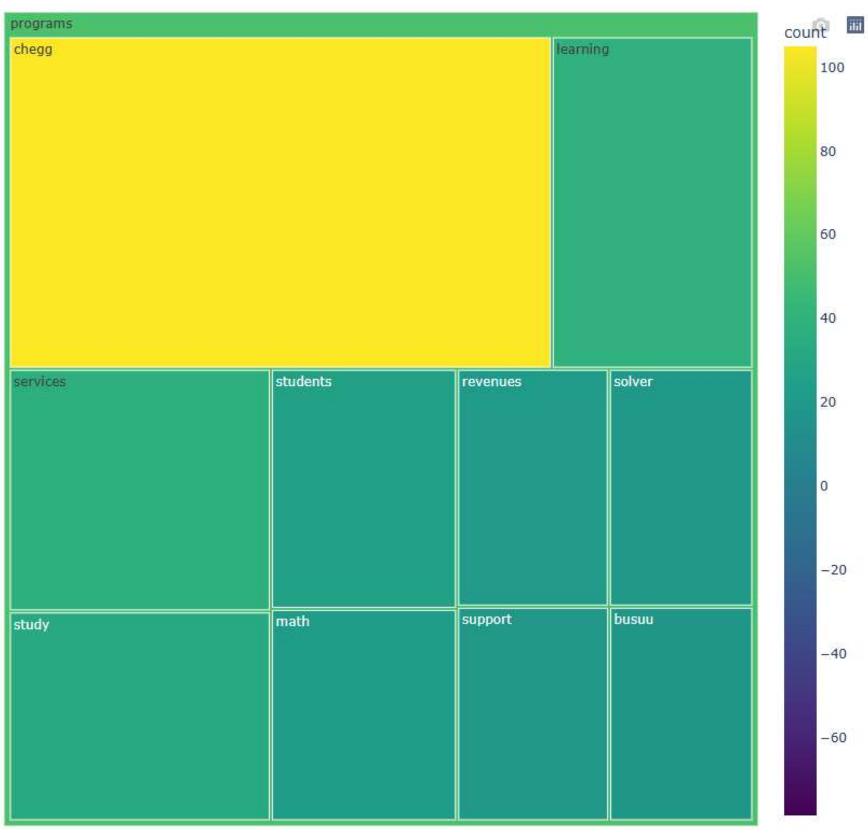
Sunburst:


ESG
students
programs
data security
legal
marketing
diversity
cost
recruiting
income
graduation rate
data management
students 92
chegg 35
student 34
learning 34
services 23
help 21
education 16
support 16
us 14
believe 13
busuu 18
support 19
solver 20
revenues 21
math 23
students 26
study 32
services 37
learning 39
chegg 105
data 39
business 32
laws 25
privacy 25
information 24
us 20
students 18
services 17
including 16
use 15
us 27
claims 22
business 19
content 15
students 14
laws 13
subject 13
states 12
intellectual 12
property 12
marketing 25
students 23
services 18
advertising 16
search 15
expenses 13
new 12
diversity 26
board 12
people 11
students 11


students
students 92
chegg 35
student 34
learning 34
services 23
help 21
education 16
support 16
us 14
believe 13

treemap:


ESG Report
programs
chegg
services
study
students
revenues
solver
learning
math
support
busuu
students
students
learning
student
services
education
support
chegg
help
us
believe
data security
data
laws
us
students
privacy
services
including
business
information
use
marketing
marketing
students
services
search
expenses
advertising
new
sales
products
use
diversity
diversity
board
people
employees
employee
chegg
inclusion
culture
focus
legal
us
business
students
subject
intellectual
claims
content
laws
property
states
cost
students
education
cost
research
services
print
less
recruiting
data
global
income
graduation rate
count
100
80
60
40
20
0
−20
−40
−60


programs
chegg
learning
services
students
revenues
solver
study
math
support
busuu
count
100
80
60
40
20
0
−20
−40
−60

Conclusion:

Using the topics_modeling, we can understand what topics are in a document and what keywords are mentioned. So we know companies talking more about which topics more in a document and what words they use to express the topic. At the same time, keyword_extractor gives us a variety of visualization options, allowing us to get a clearer and more intuitive view of keyword information.